Exploiting Document-Level Context for Data-Driven Machine Translation

Ralf D. Brown
Carnegie Mellon University Language Technologies Institute
5000 Forbes Ave
Pittsburgh, PA 15213 USA
ralf@cs.cmu.edu

Abstract

This paper presents a method for exploiting document-level similarity between the documents in the training corpus for a corpus-driven (statistical or example-based) machine translation system and the input documents it must translate. The method is simple to implement, efficient (increases the translation time of an example-based system by only a few percent), and robust (still works even when the actual document boundaries in the input text are not known). Experiments on French-English and Arabic-English showed relative gains over the same system without using document-level similarity of up to 7.4% and 5.4%, respectively, on the BLEU metric.

1 Introduction

Corpus-based machine translation systems have made considerable strides since the original Statistical MT (Brown et al., 1990; Berger et al., 1994) and Example-Based MT (Nagao, 1981; Nagao, 1984) systems, but still typically treat both the training data and the test input as isolated, entirely independent sentences. In practice, however, both the training and test data consists of a series of complete documents rather than isolated sentences.

While a number of researchers have added contextual information to the translation process in recent years, one area which has not been explored is the use of document-level context to affect the MT system's translations, This paper presents a method for adjusting phrasal translation scores in an EBMT system based on the similarity between the input document and the training documents containing matching examples.

After a brief review of related research in Section 2, the method for exploiting document level similarity is presented in Section 3, the evaluation of the method is described in Section 4, and the results of the evaluation are shown in Section 5.

2 Related Work

A number of researchers have investigated means of incorporating contextual information into corpus-driven MT systems.

For example, Brown (2005) used both intra- and inter-sentential context to affect the weighting of retrieved examples in an EBMT system (Brown, 1996). The intra-sentential context was used to boost matches which are contained within larger matches for the current input sentence, thus biasing overall translations away from the translations for matches which do not have additional context within the training instance. The inter-sentential context was used to boost matches from training instances located within a small window of a training instance which was used in the translation of the prior input sentence. While intra-sentential context proved to be consistently beneficial, inter-sentential context was helpful less often, and the combination of the two bonuses could even harm performance.

Lü *et al* (2007) applied similar sentence-specific weighting to a Statistical MT (SMT) system. In their offline version, they first select a subcorpus of sentences which are similar to the sentences in the test data or target domain, then boost the weights of selected sentences by increasing their counts within

the entire corpus (effectively appending the selected subcorpus to the training data). For their online variant, they produce several translation models using the offline variant to generate adapted corpora, then use each input sentence as an information retrieval query to determine which submodels contain similar sentences and weight each submodel according to the total number of retrieved training sentences.

Other researchers have investigated methods of subsampling the training data to generate a test set-specific corpus. One example is the work by Hildebrand *et al* (2005), which applied information retrieval techniques to select sentences similar to the test set. Such an approach substantially improves performance, but is not suitable for production systems since it requires retraining for each input file desiring adaptation.

More recently, Gimpel and Smith (2008) added within-sentence contextual features to phrase-based SMT (Koehn et al., 2003). These contextual features are added to the source-language side of the phrase table, allowing for better prediction of translations without major modifications to the decoder (in fact, Gimpel and Smith were able to avoid modifications entirely by appending a unique identifier to each token of input prior to computing the phrase table). As with Brown's intra-sentential context, only the current sentence is considered.

Statistical MT systems are also becoming more EBMT-like, performing on-the-fly lookups or phrase-table generation rather than using a static phrase table. Examples include Vogel's (2005) "online" system and Carpuat and Wu's (2007) system incorporating phrase-sense disambiguation. Other systems use subsampling or related techniques to generate document- or test set-specific phrase tables; (Gimpel and Smith, 2008) generate such a test set-specific phrase table as part of applying source context.

3 Document-Level Context

The basic idea behind exploiting document-level context is to most heavily weight examples retrieved from training documents which are most similar to the input document. When the final translation candidates are computed from a weighted combination of retrieved examples, the scores will be biased toward translations coming from the more similar training documents. The intuition here is that a similar training document is more likely to use the same word senses than a dissimilar document.

Exploiting document-level context begins during training, but the requirements at training time are trivial: the corpus must be marked up with original document information and the system needs to record which sentence pairs belong to each training document.

At translation time, explicit document begin and end markers are used to permit multiple documents in a single data file. Upon encountering a begin-document marker, the system begins accumulating statistics and storing the input sentences. When the matching end-document marker is reached, each training document is assigned a similarity score based on the accumulated statistics and the stored sentences are then translated.

To simplify implementation, the existing matching code in our EBMT system is used as the basis for computing similarity scores. Thus, in the first pass over the input document, corpus matches are found by consulting the corpus index just as they are during normal translation, but the matches are not actually retrieved. Instead, the match records are filtered to eliminate the most frequent, and least indicative, n-grams; specifically, those occurring often enough in the complete corpus to invoke subsampling (as described below). The remaining low-frequency n-gram records are then processed to determine which training documents contain the matches, and the count for each document is incremented by the total number of words matched by the filtered records. Once all sentences in the input document have been processed, the similarity score is computed as the normalized (by document length) and scaled count, such that the document with the highest proportion of matched words receives a score of 1.0 and any documents with no matches at all receive a score of 0.0. Figure 1 contains a pseudo-code description of this process.

In addition to simplicity of implementation (due to the re-use of existing mechanisms), the n-gram match similarity score has the additional advantage that it does not require additional storage as would term vectors for cosine similarity, message hashes, or alternative representations such as

Initialize match count for each training document to zero
For each sentence in input document:

1. find all n-gram matches in training corpus
2. eliminate all matches for source phrases which occur more than $maxdups$ times in the entire corpus
3. for each remaining n-gram match, increment the match count of the document containing it by the length of the match

Normalize counts into similarity score:

1. divide match count for each training document by its length in sentences
2. normalize to [0.0,1.0] by dividing each document's score by the highest score

Assign each training document one of 128 ranks based on its similarity score

Figure 1: Similarity-scoring procedure

For each input sentence

1. find all n-grams in the training corpus matching any part of the input sentence
2. subsample the instances of any n-grams occurring more than $maxdups$ times in the corpus, using in order of priority
   (a) complete matches of training instances
   (b) matches with the most context words in the training instance
   (c) **matches from documents with the highest similarity rank**
   (d) uniform sampling within whichever of the above categories which would exceed $maxdups$
3. for each remaining match
   (a) retrieve the sentence pair
   (b) perform a subsentential alignment to determine the candidate translation
   (c) assign a score and a weight to the resulting translation pair, using
      - local context bonus
      - optional source-specific weight
      - optional corpus position-specific weight
      - **document similarity score for the training instance**
4. perform a weighted merge of all equivalent translation pairs
5. if necessary, limit the number of candidate translations for a source-language n-gram to the k top-scoring alternatives
6. place the final translation candidates in a lattice and apply a stack decoder to determine the best overall translation

Figure 2: Translation process in pseudo-code

locally-weighted bags of words.

The EBMT system used for the experiments described in the next section functions in the same manner as those of Brown (2005) and Phillips *et al* (Phillips and Cavalli-Sforza, 2006; Phillips et al., 2007). See Figure 2 for pseudo-code; the boldfaced portions indicate the modifications to support document similarity.

Our method modifies steps 2 and 3(c) in the following ways. In Step 2, the baseline system orders n-gram matches by the number of additional words which match in that training example (the amount of available local context) and selects instances in decreasing order of context until a particular level of context would result in more than $maxdups$ instances. This final level is then uniformly subsampled to produces exactly $maxdups$ total instances. For the enhanced system, the training documents are ranked by similarity value and the ranks are quantized into 128 levels. The quantized ranks are then used as a tie-breaker in the final level of context selection, and uniform subsampling is only used among the instances in the document rank which would result in more than $maxdups$ instances.

In Step 3(c), the baseline system assigns a static weight to the translation pair. This weight is computed as a combination of the local context bonus (as in (Brown, 2005)) and optional source-specific or corpus position-specific weights. The optional weights were not used in this work except for a single run described in Section 6; the local context bonus is still helpful even with context-biased subsampling because a large proportion of all n-grams occur infrequently enough that subsampling is not applied. For the enhanced system, an additional factor based on the similarity measure is multiplied with the baseline weight. This additional factor is

$$(1 - \lambda) + \lambda \times similarity$$

where λ is a tuneable parameter used to set the strength of the bias toward the most similar training documents. Translation performance is typically maximized for λ values of 0.75–0.83.

The computational cost of the changes to steps 2 and 3(c) is negligible. In both cases, the additional work is essentially a hash-table lookup followed by a few arithmetic operations. The bulk of the overhead in applying document similarity lies in actually computing the similarity scores. Fortunately, index lookups are very quick compared to actually retrieving, aligning, and merging the matches found by the index lookup. The cost of the additional index lookups for each sentence and accumulation of match counts is only a few percent of the total run-time, and is often overshadowed by changes in run-time resulting from differences in the decoder's search caused by altered scores (runs using document boundaries are occasionally marginally *faster* than those without).

4 Evaluation

4.1 Data Sets

Two different training corpora were used, one for Arabic-English and the other for French-English. Multiple test sets were translated for each of these language pairs.

For Arabic-English, a subset of the training data permitted for the 2008 NIST Open Machine Translation evaluation (MT08) constrained-data track was used. This subset consisted of all permissible newswire data plus a small portion of the United Nations corpus, excluding sentence pairs where source and target lengths differed by more than a factor of four or the source contained more than 255 tokens. The Arabic text was converted from UTF-8 encoding to the Buckwalter latinization and several common affixes were separated into standalone tokens. This provided a total of 1.4 million sentence pairs (46.2 million source tokens) of training data. The language model for the decoder was built from the target half of the parallel corpus plus the Xinhua portion of the English Gigaword corpus (Graff et al., 2007), Third Edition (LDC2007T07).

To tune parameters, two subsets of the 2003 MT Evaluation (MT03) test set were used, a 96-sentence set for the actual tuning process and a 220-sentence validation set to determine which of two tuning results for each of the test conditions to use on the blind test sets. The blind test sets were the 2004 and 2005 NIST MT evaluation test sets (MT04 and MT05) and the "NIST" portion of the 2006 NIST Open MT evaluation (MT06-NIST). All test sets were preprocessed in the same manner as the training data. Each of these data sets has at least four reference translations (five for MT05). Three versions of the test sets were prepared: one with document boundaries preserved in the form of begin-document and end-document markers, one with document boundaries removed, and one with document boundaries arbitrarily placed around each group of eight sentences.

Using the boundary-less test files as input forces the EBMT system to revert to baseline operation

even when trained on a corpus annotated with document boundaries, since by default every training document is scored equally.

For French-English, the training data was the Europarl corpus (Koehn, 2005), version 3 [1]. All sentence pairs of the designated training portion were used, except those where the source and target differed by more than a factor of 2.5 in length and the shorter of the two was at least eight tokens in length, or the source sentence exceeded 255 tokens. This provided a total of 1.3 million sentence pairs (43.5 million source tokens) of training data. The language model for the decoder was trained solely on the target half of the parallel corpus.

To tune parameters, a 40-sentence subset of the "devtest2006.fr" file was used, consisting of eight contiguous five-sentence groups each marked as a document for the context-sensitive runs. Parameters were tuned separately for the baseline and context-sensitive conditions using an adaptive grid-based coordinate search over some 40 parameters including $maxdups$, beam width, maximum number of alternative translations to place in the lattice, standard decoder features (length ratio, language model score, translation score), and when appropriate the strength of the document-similarity bias. The blind test sets were the French "test2006" and "test2007" sets, which were originally evaluation data and became development test sets for the WMT08 evaluation. Each of these data sets has a single reference translation. Two versions of the test sets were prepared: one without document boundaries and one with document boundaries arbitrarily placed every five sentences ("test2006") or every ten sentences ("test2007") since the original document boundaries were not available.

4.2 System Training

For the experiments described below, the EBMT system was trained as a straight-forward string-matching EBMT system, without generalized matching through clustering such as in (Brown, 2000) or structural matching such as in (Phillips et al., 2007). GIZA++ (Al-Onaizan et al., 1999) word alignments were used for the Arabic-English data to drive the subsential aligner; for French-English, we used a much faster heuristic approach driven by a bilingual lexicon generated with an algorithm that roughly corresponds to an amalgamation of the GIZA++ IBM Model 1 and HMM phases with the competitive linking algorithm of (Melamed, 1997).

[1] Available at `http://www.statmt.org/europarl/`

4.3 Evaluation Metrics

Performance was evaluated using the `mteval-v11b.pl` script made available by NIST. This script reports both the BLEU metric (Papineni et al., 2002) and the variant thereof proposed by George Doddington at NIST (Doddington, 2002). Parameters were tuned to maximize the value of the BLEU metric.

BLEU measures the proportion of n-gram overlap between the MT system's output and one or more human reference translations. The NIST metric modifies the n-gram matching to assign different weights to individual n-grams depending on their information content.

The METEOR metric (Banerjee and Lavie, 2005) has not yet been integrated into our workflow, but that integration is planned for the near future.

The statistical significance of the results was tested with the Wilcoxon Signed-Rank Test (Wilcoxon, 1945), a nonparametric alternative to the paired t-test. Unlike the t-test, the Signed-Rank test does not assume normal distribution of values in the population, nor does it assume that the scale of measurement is an equal-interval scale. To generate the values used for the Signed-Rank test, the test outputs were uniformly split into either 10 or 20 segments, depending on the total number of sentences in the test set, and each segment was individually scored with the `mteval` script. Bootstrapping as described by (Zhang et al., 2004) was not used because the test sets are sufficiently large to provide enough truly independent samples for statistical significance tests.

5 Results

Tables 1 and 2 present the results of runs for Arabic-to-English and French-to-English, respectively. For each data set, the tables show the NIST and BLEU scores of the baseline system (no document boundaries in the input text), the enhanced system with arbitrary uniformly-spaced document boundaries, and

	BLEU scores		
Data Set	Baseline	True Docmuent Boundaries	8-sentence Boundaries
MT04	0.37726	0.39787 (+0.02061/+5.4%)	0.39775 (+0.02049/+5.4%)
MT05	0.44043	0.45205 (+0.01162/+2.6%)	0.45489 (+0.01446/+3.2%)
MT06	0.33390	0.33482 (+0.00092/+0.3%)	0.33590 (+0.00108/+0.6%)
	NIST scores		
Data Set	Baseline	True Document Boundaries	8-sentence Boundaries
MT04	9.0727	9.6130 (+5.9%)	9.5839 (+5.6%)
MT05	9.9539	10.2173 (+2.6%)	10.2464 (+2.9%)
MT06	8.8569	8.8681 (+0.1%)	8.8828 (+0.3%)

Table 1: Performance of Arabic-English translations, tuned on MT03

Data Set	Condition	BLEU score	NIST score
2006	No docs	0.24896	6.7063
2006	Doc-5	0.26614 (+0.01718/+6.9%)	6.9150 (+3.1%)
2007	No docs	0.25218	6.7719
2007	Doc-10	0.27088 (+0.01860/+7.4%)	7.0100 (+3.5%)

Table 2: Performance of French-English translations

if appropriate the enhanced system with true document boundaries for the input text. In the case of uniform document boundaries, the number of sentences per pseudo-document (5, 8, or 10) is indicated.

It can be seen from Table 1 that performance gains for Arabic are quite uniform between the two metrics and between the true-boundaries and arbitrary-boundaries cases. The MT06 data set is known to be "harder" than earlier MT0x sets, hence the overall lower scores.

Table 2 shows consistent improvements across the two test sets, with BLEU gains about twice as large as NIST gains. Unlike the Arabic case, both test sets are from the same epoch and are thus more similar to each other than the MT0x test sets are.

In addition, one quick experiment was performed on the French-English data to ascertain how much of the improvement is due to the alteration of the subsampling of highly-frequent n-grams. This was done by simply setting the value of λ to zero, allowing the subsampling to take advantage of the document similarity scores but not using them to reweight candidates. On the "test2006" set, this resulted in a BLEU score of 0.24928 (+0.1% compared to the baseline) and NIST score of 6.6946 (-0.1%). As expected, the use of document similarity as a tie-breaker in the sampling has only a very small effect.

The difference between the baseline and true-document case is statistically significant for MT04, as is the difference between baseline and arbitrary 8-sentence boundaries. For MT05, the situation is different, and somewhat counter-intuitive: the difference between true boundaries and 8-sentence boundaries is statistically significant (p=0.0235), while the much larger difference between baseline and true boundaries is not (p=0.1736). This is the result of the true-document case peforming worse than the baseline on the portion of the test set originating from the Xinhua news service while performing much better than the baseline on the AFP portion of the test set. In contrast, the eight-sentence condition showed a small but much more uniform improvement over the true-boundary condition. For MT06, none of the differences are statistically significant (p$\geq$0.117).

On the French-English data, the improvements for both 2006 and 2007 are highly significant

($p<0.0001$).

The excellent performance of the uniform-boundaries test case for Arabic-English came as somewhat of a surprise, since a large fraction of the segments contain text from two separate test documents (the average length of documents in MT04/05/06 is 6.8, 10.5, and 17.3 sentences, respectively). We had expected a substantially smaller improvement due to less-accurate similarity scoring.

6 Further Analysis

After the initial submission of this paper, further experiments were performed to characterize the reason for the very small improvement on MT06.

First, the system was tuned on subsets of MT05 and MT06, and translation performance compared to the system tuned on MT03. For MT05, the first six and last six documents of the test set, totalling 131 sentences, were selected as a tuning set; for MT06, random documents totalling 127 sentences were selected as a tuning set. Table 3 shows the results of this experiment, as well as a further experiment in which the parameters tuned on the MT06 baseline condition were used for both baseline and document-boundary cases. Naturally, the cases of MT05 tuned with MT05 and MT06 tuned with MT06 are not fully comparable to the other cases since a small portion of the test sets were used for tuning, so the test is no longer blind. However, analyzing the results in those cases is still instructive.

Consistent with the hypothesis that decreased improvement in the later MT0x test sets is due to increased differences between those test sets and MT03, performance on MT04 and MT05 is better when tuned with either MT05 or MT06 than when tuned with MT03. Likewise, decreased performance for MT06 tuned on MT05 lends support to the conjecture that MT06 is in some fundamental way different from the other MT0x data sets, in a manner which decreases the effectiveness of document-similarity weighting.

Significance testing with the Wilcoxon Signed-Rank test shows that MT04 tuned on MT05 has a statistically significant improvement over the baseline (p=0.0054), as does MT04 using only the MT06 baseline parameters (p=0.030). MT05 showed a significant improvement whether tuned on MT05 or MT06, or using only the MT06 baseline parameters (p=0.0074, 0.0054, and 0.0178, respectively). MT06 tuned on MT05 has a statistically significant improvement (p=0.0012), as does MT06 using only the parameters tuned for the baseline case (p=0.034), while MT06 using normal tuning on MT06 does not produce a significant change (p=0.529). Of note is that all three test sets showed a significant improvement from adding the document-similarity weighting when run using MT06 parameters tuned without that weighting.

Examining the actual values of the tunable parameters produced during tuning shows one major outlier relevant to the document-similarity weighting. For each of the tuning sets, `Doc-Sim-Weight`, the λ value determining the strength of the similarity preference, was tuned to values ranging from 0.804 to 0.832; for every tuning condition *except* the baseline case on MT06, $maxdups$ received values ranging from 600 to 615. In contrast, the MT06 baseline tuning produced a value of 225.

This finding led to the conjecture that a reduced number of corpus matches overall for the MT06 test set was allowing a larger proportion of matches from UN training documents to be used. As the UN documents are quite different from newswire documents, they could adversely influence the translation candidates produced by the EBMT engine. To test this conjecture, MT06 was translated without document boundaries using the tuned parameters for the no-boundaries (baseline) condition but reducing the weight of the UN training documents to 0.1 relative to all other documents. Doing so increases the BLEU score from 0.33799 to 0.33929, indicating that UN text is indeed adversely affecting the translation of the MT06 test set.

7 Conclusion

This paper has presented a simple enhancement to an existing EBMT system that efficiently and robustly provides substantial gains from exploiting document-level similarity between the training corpus and the input being translated. This enhancement, or a variation thereof, can also be added to statistical MT systems which make use of dynamic phrase tables generated at runtime.

Tuning Set	Test Set BLEU scores Baseline/True-Boundaries		
	MT04	MT05	MT06
MT03	0.37726/0.39787	0.44043/0.45205	0.33390/0.33482
MT05	0.38742/0.40664	0.45261/0.46757	0.31388/0.32694
MT06	0.41421/0.40674	0.44131/0.45885	0.33799/0.33541
MT06-nodoc	0.41421/0.42120	0.44131/0.44757	0.33799/0.34302

Table 3: Performance comparison of Arabic-English translations

8 Future Work

This work is still in its early stages, and there are a number of directions left to investigate.

Are there better similarity measures? The existing one was chosen for ease of implementation and because it does not require extra storage, but other methods such as cosine similarity over term vectors, message hashes/digests such as Nilsimsa (Damiani et al., 2004; Nilsimsa, 2003), or alternative representations such as locally-weighted bags of words (Lebanon et al., 2007) may be more effective.

Is it possible to subdivide documents in a better manner? Particulary for parliamentary proceedings such as the Europarl and UN corpora, the documents are typically long and may contain multiple differing sections. Automatically determining additional boundaries based on changing content should permit finer-grained and thus more accurate weighting of the training documents. Techniques developed during the DARPA Topic Detection and Tracking project (Allan et al., 1998; Carbonell et al., 1999) for the Story Segmentation task are likely to be useful.

Is there a better way to incorporate the document similarity score into the overall score for a translation candidate? The present linearly-damped multiplicative factor was merely the first approach attempted.

Why did performance deteriorate on the MT05 Xinhua data when it improved on all other data? How did uniform pseudo-boundaries manage to outperform true boundaries? These two conditions may be artifacts of overfitting the tuning set, but the questions still need to be answered through additional analysis.

In addition, a variation of the Gimpel and Smith (Gimpel and Smith, 2008) approach to within-sentence context is being implemented by further modifying the match filtering and weighting steps to take easily-computable features such as location within the source-language sentence into account.

9 Acknowledgements

The author would like to thank Aaron Phillips for providing a pre-processed version of the MT08 data and the anonymous reviewers for suggestions which improved the final version of this paper.

References

Yaser Al-Onaizan, J. Curin, M. Jahr, Kevin Knight, John Lafferty, I.Dan Melamed, Franz Joseph Och, D. Purdy, Noah A. Smith, and David Yarowsky. 1999. Statistical Machine Translation: Final Report. In *Proceedings of the Summer Workshop on Language Engineering*. John Hopkins University Center for Language and Speech Processing.

James Allan, Jaime G. Carbonell, George Doddington, Jon Yamron, and Yiming Yang. 1998. Topic Detection and Tracking Pilot Study Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understranding Workshop*, Feb.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *Porceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, June.

Adam L. Berger, Peter F. Brown, Stephen A Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, and Luboi Urei. 1994. The Candide System for Machine Translation. In *Proceedings of the ARPA Human Language Technology Workshop*.

Peter Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin.

1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16:79–85.

Ralf D. Brown. 1996. Example-Based Machine Translation in the PANGLOSS System. In *Proceedings of the Sixteenth International Conference on Computational Linguistics*, pages 169–174, Copenhagen, Denmark. http://www.cs.cmu.edu/~ralf/papers.html.

Ralf D. Brown. 2000. Automated Generalization of Translation Examples. In *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*, pages 125–131.

Ralf D. Brown. 2005. Context-Sensitive Retrieval for Example-Based Machine Translation. In *Proceedings of Workshop: Example-Based Machine Translation, The Tenth Machine Translation Summit*, pages 12–16, September. http://www.cs.cmu.edu/~ralf/papers.html.

Jaime Carbonell, Yiming Yang, John Lafferty, Ralf D.Brown, Tom Pierce, and Xin Liu. 1999. CMU report on TDT-2: Segmentation, Detection and Tracking. In *Proceedings of the DARPA Broadcast News Workshop*, pages 117–120, San Francisco, CA. Morgan Kaufmann Publishers, Inc.

Marine Carpuat and Dekai Wu. 2007. How Phrase Sense Disambiguation Outperforms Word Sense Disambiguation for Statistical Machine Translation. In *Proceedings of the Eleventh Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, pages 43–52, Sep.

E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati. 2004. An Open Digest-based Technique for Spam Detection. In *Proceedings of the 2004 International Workshop on Security in Parallel and Distributed Systems*, September. http://seclab.dti.unimi.it/Papers/pdcs04.pdf.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using n-gram Cooccurrence Statistics. In *Proceedings of the Human Language Technology Conference (HLT)*, pages 128–132.

Kevin Gimpel and Noah A. Smith. 2008. Rich Source-Side Context for Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 9–17, June.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. English Gigaword Third Edition. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T07.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation base on Information Retrieval. In *Proceedings of the 10th EAMT Conference "Practical Applications of Machine Translation" (EAMT 2005)*, pages 133–142, May. http://www.mt-archive.info/EAMT-2005-Hildebrand.pdf.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL*, pages 127–133.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summix X)*, pages 79–86.

Guy Lebanon, Y. Mao, and J. Dillon. 2007. The Locally Weighted Bag of Words Framework for Document Representation. *Journal of Machine Learning Research*, 8:2405–2441, Oct.

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL-2007)*, pages 343–350, June. http://acl.ldc.upenn.edu/D/D07/D02-1036.pdf.

I. Dan Melamed. 1997. A word-to-word model of translational equivalence. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 490–497, Somerset, New Jersey. Association for Computational Linguistics.

Makoto Nagao. 1981. . In *Proceedings of the International NATO Symposium*. NATO Publications, October.

Makoto Nagao. 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In Alick Elithorn and Ranan Banerji, editors, *Artificial and Human Intelligence*, pages 173–180. North-Holland.

Nilsimsa. 2003. Nilsimsa message digest. http://ixazon.dynip.com/~cmeclax/nilsimsa.html (last visited 06dec2007).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July. http://acl.ldc.upenn.edu/P/P02/.

Aaron B. Phillips and Violetta Cavalli-Sforza. 2006. Arabic-to-English Example Based Machine Translation Using Context-Insensitive Morphological Analysis. In *Journées d'Etudes sur le Traitement Automatique de la Langue Arabe (JETALA)*, June.

Aaron B. Phillips, Violetta Cavalli-Sforza, and Ralf Brown. 2007. Improving Example Based Machine Translation Through Morphological Generalization

and Adaptation. In *Proceedings of the Eleventh Machine Translation Summit (MT Summit XI)*, September.

Stephan Vogel. 2005. PESA: Phrase Pair Extraction as Sentence Splitting. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*.

F. Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics*, 1:80–83. online tool: http://-faculty.vassar.edu/lowry/wilcoxon.html.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International COnference on Language Resources and Evaluation (LREC-2004)*, pages 2051–2054. `http://www.mt-archive.-info/LREC-2004-Zhang.pdf`.